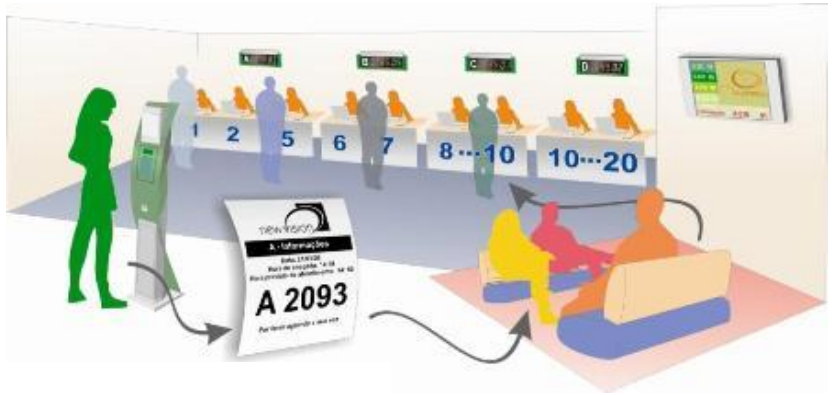
Filas de Espera



Filas de Espera

A **teoria das filas de espera** (ou *queueing theory*) é um ramo da matemática aplicada que estuda **sistemas onde chegam pedidos, tarefas ou pessoas que precisam de aguardar por um serviço**. Analisa o comportamento das filas para prever tempos de espera, capacidade necessária, desempenho e eficiência.

Em termos simples, serve para **modelar situações onde a procura por um serviço é variável, mas os recursos para atendê-la são limitados**.

Filas de Espera - Para que se utilizam?

Otimizar o desempenho

- Avalia o tempo de espera médio, a taxa de atendimento e a ocupação de servidores, ajudando a melhorar a eficiência de sistemas digitais ou físicos.

Dimensionar os recursos

- Ajuda a decidir quantos servidores, máquinas, instâncias ou processos são necessários para atender a uma determinada carga de solicitações.

Filas de Espera - Para que se utilizam?

Reduzir custos

- Evita tanto a subutilização (muitos recursos ociosos) quanto a sobrecarga (esperas longas e falhas).

Prever comportamentos sob diferentes cargas

- Permite simular o impacto de picos de uso, interrupções, latência e aumento de utilizadores.

Garantir a qualidade de serviço (QoS)

- Muitas aplicações dependem de limites de tempo de resposta e as filas ajudam a prever se esses limites serão cumpridos.

Servidores Web

- Quando um servidor recebe múltiplos pedidos HTTP simultâneos, estes entram numa fila de processamento. A teoria das filas ajuda a determinar:
  - quantas threads ou máquinas o servidor deve ter,
  - como o equilíbrio de trabalho deve funcionar,
  - quando a fila ficará saturada sob picos de tráfego.

Sistemas distribuídos e computação na nuvem

- Plataformas como AWS, Azure ou Google Cloud utilizam modelos de filas para:
  - autoscaling (subir/descer nº de máquinas com base no trabalho),
  - gestão de filas SQS, RabbitMQ, Kafka,
  - análise de latência e throughput.

Bases de dados

- Operações concorrentes entram em filas de espera por bloqueios (*locks*) ou transações.
  A teoria das filas ajuda a prever:
  - tempo de espera por recursos concorrenciais,
  - impacto do *locking*,
  - tempo médio de transação.

Processadores e sistemas operativos

- Numa CPU, os processos competem por:
  - ciclos do processador,
  - acesso à memória,
  - Input/Output.

Os componentes do **sistema operativo/schedulers** como Round Robin, SJF, FIFO são baseados em modelos de filas.

Redes de computadores

- As bases de dados acumulam-se em filas em:
  - routers,
  - switches,
  - buffers de rede.
- A teoria das filas permite prever:
  - Latência do trabalho,
  - perda de bases de dados,
  - Congestão de trabalho.

Resumo final

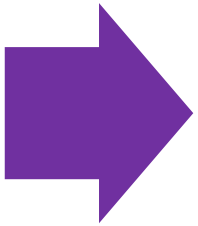
- A teoria das filas permite:
  - **modelar** sistemas com concorrência,
  - **otimizar** desempenho,
  - **prever** congestionamentos,
  - **dimensionar** recursos,
  - **melhorar** a experiência de utilizadores.

É uma ferramenta essencial em engenharia informática, especialmente em sistemas de alto desempenho, redes, cloud computing e servidores web.

Filas de Espera

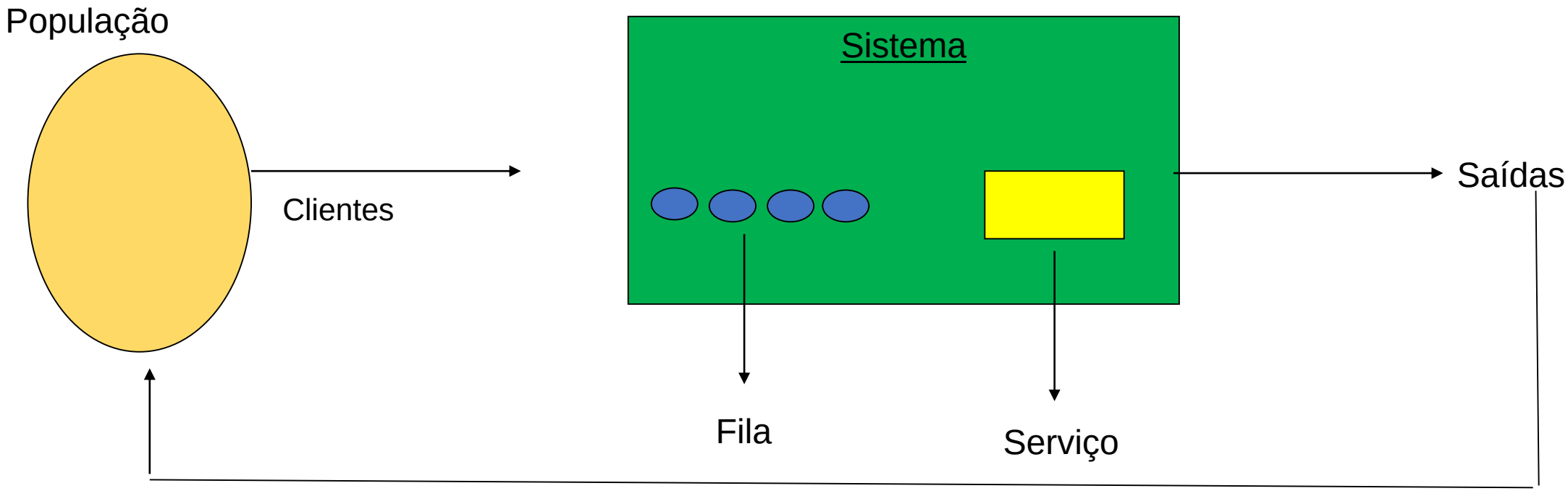
Uma fila de espera é um fenómeno do dia-a-dia, no qual existem clientes e serviços

Clientes: pessoas, veículos ou outras entidades físicas ou conceptuais que necessitam de um



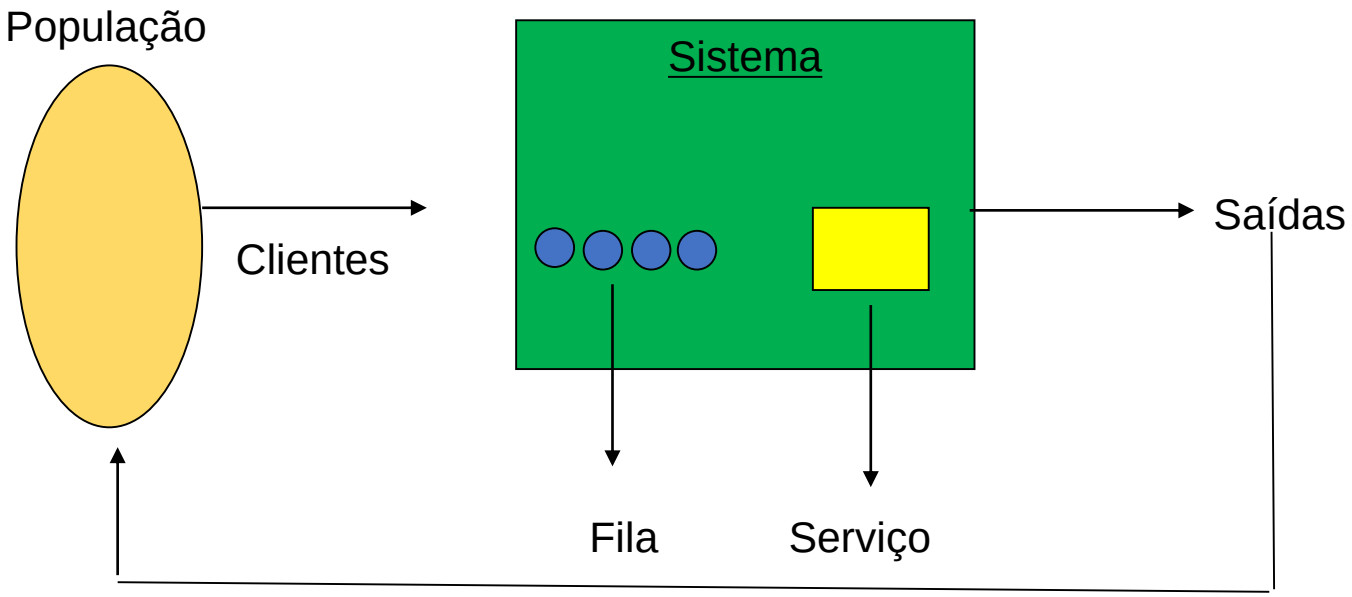
Serviço: pelo qual podem ter que esperar numa fila física ou conceptual (a fila pode não existir conceptualmente, por exemplo senhas com número de atendimento).

Estrutura de um sistema de fila de espera



Fonte ou população, que gera os clientes que vão chegar ao sistema. Exemplo: a urgência de um hospital. A população neste caso é constituída por todos os habitantes que pertencem à zona de influência do hospital.

Estrutura de um sistema de fila de espera



Fila + Serviço = Sistema

Número de clientes no sistema (em cada instante) = **estado do sistema (*n*)**.

Fila, constituída pelos clientes à espera de serem atendidos (não inclui o(s) cliente(s) em atendimento).

Serviço ou atendimento, que pode ser constituído por um ou mais postos de atendimento.

Filas de espera: conceitos básicos

Muitos modelos clássicos de filas de espera têm a mesma base matemática dos modelos estatísticos de vida e morte.

Porque existe esta relação?

Tanto os **modelos de filas de espera** como os **modelos de vida e morte** utilizam processos estocásticos do tipo **Markoviano**, onde o estado do sistema muda ao longo do tempo devido a: **entradas** (chegadas / nascimentos) e **saídas** (partidas / mortes).

Assim, embora estudem fenómenos diferentes, **a estrutura matemática é muito parecida**.

Um **processo de Markov** é um tipo de modelo matemático usado para descrever sistemas que evoluem ao longo do tempo, passando de um estado para outro **de forma aleatória**, mas com uma característica muito especial: o **futuro depende apenas do estado atual não importando como se chegou até ali**.

Ou seja, **só o momento presente importa** para determinar o que pode acontecer a seguir. “Para prever o próximo passo, basta saber onde estou agora, e não o caminho que fiz para chegar aqui.”

Modelo de vida e morte : conceitos básicos

Como funciona de forma muito simples?
Imagine uma população qualquer: pessoas, animais, bactérias… ou até algo mais abstrato como tarefas num sistema.

Num dado momento:
Se acontece um **nascimento**, a população passa de *N* para *N + 1*.
Se acontece uma **morte**, a população passa de *N* para *N – 1*.

O que o modelo faz é estudar **as probabilidades** destes eventos e prever:

- se a população tende a crescer,
- se tende a desaparecer,
- ou se fica mais ou menos estável ao longo do tempo.

Modelo de vida e morte : conceitos básicos

Exemplo muito simples

Imagine que está a jogar um jogo de tabuleiro e, a cada turno, lança um dado: se o peão está na casa 5, a próxima posição depende **apenas** da casa 5 e do resultado do dado. Não importa se antes estava na casa 2, 3 ou 4. Isto é um comportamento do tipo **Markov**.

Um processo de Markov inclui:

Estados — posições possíveis do sistema (ex.: número de pessoas numa fila).

Transições — mudanças entre estados (ex.: chega alguém → estado +1).

Probabilidades de transição — a probabilidade de ir de um estado para outro.

Distribuição das chegadas

O padrão das chegadas pode ser descrito pelo tempo entre duas chegadas consecutivas (distribuição das chegadas - exponencial) ou pelo número de chegadas por unidade de tempo (distribuição do nº de chegadas - Poisson).

- ***constante:*** intervalos de tempo, entre chegadas sucessivas, fixos Ex: filas de montagem industriais;
- ***aleatório:*** os intervalos de tempo entre chegadas sucessivas não podem ser previstos com certeza → distribuições de probabilidade.

Número médio de clientes que procuram o serviço por unidade de tempo: taxa de chegada (**λ**)

- ***independente*** do estado do sistema (**λ)**;
- ***dependente*** do estado do sistema: λ_n, em que *n* é o número de clientes no sistema.

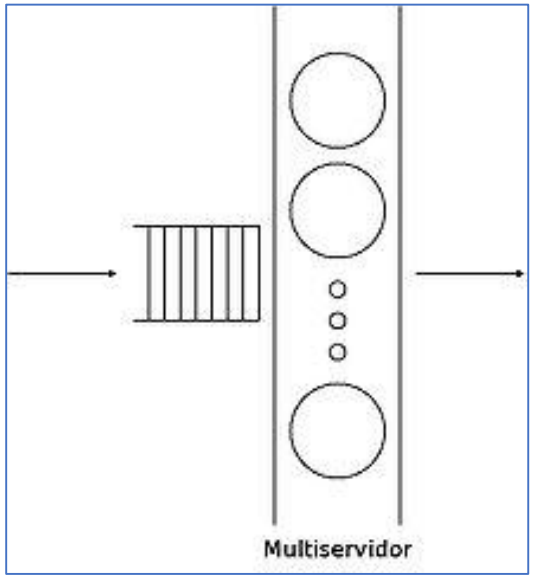
Número de filas

Fila simples: uma única fila mesmo que o servidor tenha vários postos de atendimento;

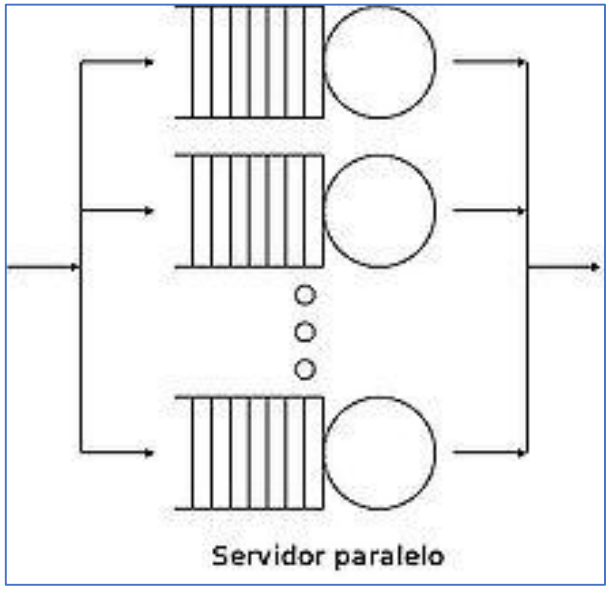
Fila múltipla: uma fila por posto de atendimento

→ cada conjunto fila/posto de atendimento constitui um sistema separado de fila de espera.

Múltiplo servidor, com fila única



Servidores paralelos



Filas de Espera – Serviço: características

Dimensão do Serviço

- Simples (máquina de aquisição de senhas; máquina multibanco)
- Grupo (por exemplo, um elevador, que atende viários clientes em simultâneo).

Distribuição do tempo de serviço

- Constante
- Aleatório: Distribuição exponencial negativa, Erlang, outra.

Filas de Espera – Serviço: características

Taxa de serviço - μ

É o número médio de clientes que podem ser atendidos por cada servidor e por unidade de tempo.

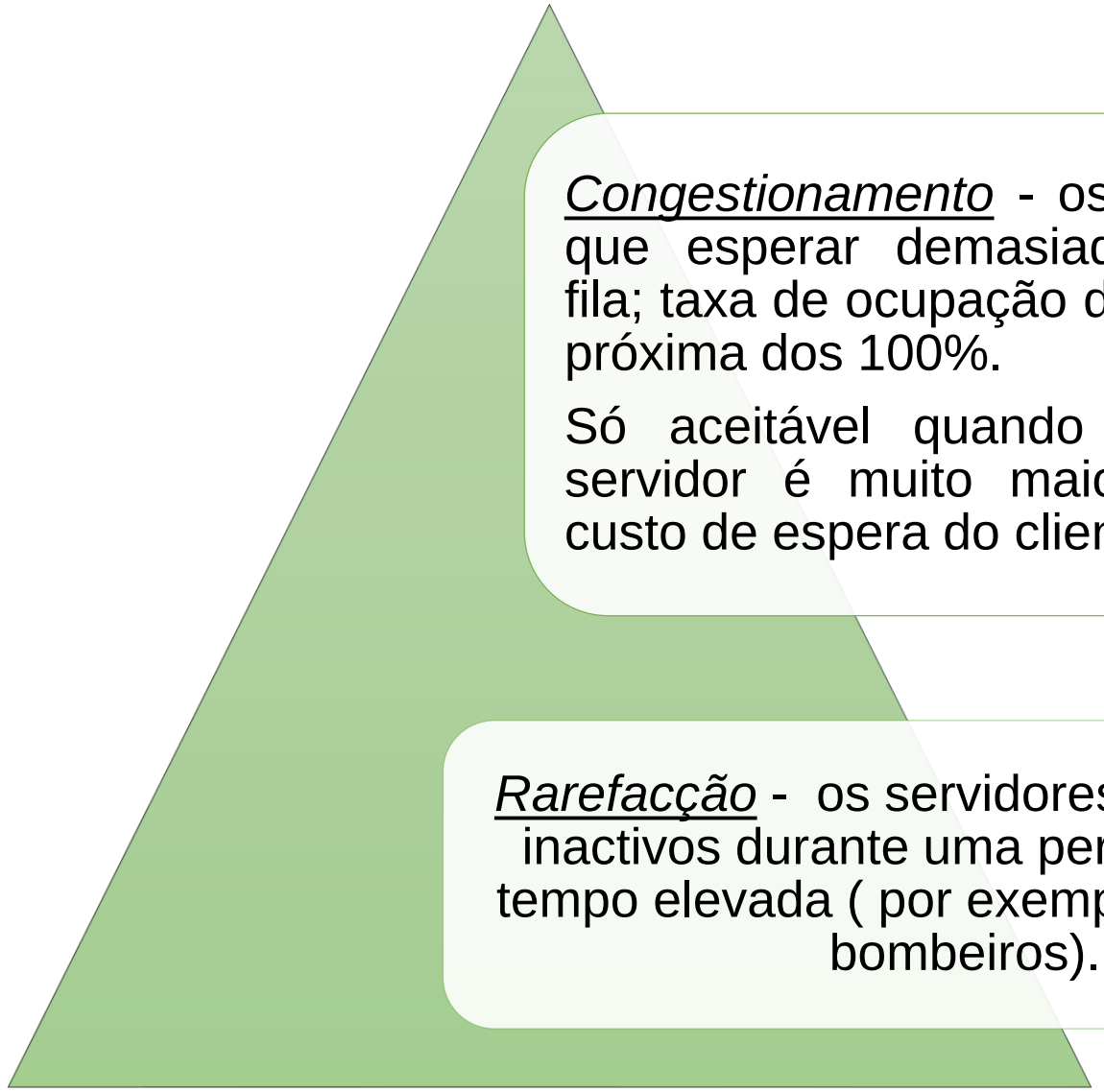
=duração média do serviço

- **Dependente** do estado do sistema (μ_n em que n é o número de clientes no sistema)
- **Independente** do estado do sistema.

É usual considerar o número médio de clientes efectivamente atendidos ≤ μ, pois pode haver momentos de inactividade.

Modelação de sistemas de filas de espera: objetivo

Evitar



Congestionamento - os clientes têm que esperar demasiado tempo na fila; taxa de ocupação dos servidores próxima dos 100%.

Só aceitável quando o custo do servidor é muito maior do que o custo de espera do cliente.

Rarefacção - os servidores permanecem inactivos durante uma percentagem de tempo elevada (por exemplo serviços de bombeiros).

Medidas de desempenho

Sigla	Descrição	Sigla	Descrição
L_q	Comprimento médio da fila	$P(W_q = 0)$	probabilidade de o tempo de espera na fila ser nulo
L	Número médio de clientes no sistema	$P(W_q > t)$	probabilidade de Wq, o tempo de espera na fila, exceder t
W_q	Tempo médio de espera na fila	$P(W > t)$	probabilidade de W, o tempo gasto no sistema, exceder t
W	Tempo médio de espera no sistema	$\frac{1}{\lambda}$	Intervalo médio entre duas chegadas sucessivas
P_n - probabilidade de existirem n clientes no sistema	o sistema estar no estado n	**S**	Nº servidores
$P(n \geq k) = \sum_{n=k}^{\infty} P_n$	probabilidade de existirem n ou mais clientes no sistema	**ρ**	Parâmetro auxiliar, designado por taxa de ocupação

Relações fundamentais de filas de espera

Relacionam o número de elementos no sistema (**L**), ou na fila (**Lq**), com os correspondentes tempos de espera **W** e **Wq**, para filas de espera em equilíbrio.

Admitindo taxas de chegada λ e de serviço μ **constantes e independentes** do estado do sistema:

$\mathbf{L = \lambda W}$	Esta relação analítica entre L e W representa o número médio de elementos no sistema
$\mathbf{L_q = \lambda W_q}$	Esta relação representa o número médio de elementos na fila
$W = W_q + \frac{1}{\mu}$	Esta expressão relaciona o tempo médio de permanência no sistema com o tempo médio de espera na fila
$L = L_q + \frac{\lambda}{\mu}$	Esta expressão relaciona o número médio de clientes no sistema com o número médio de clientes na fila

Classificação das Filas de espera – X/Y/Z/W (segundo Kendal)

X, **Y** - Distribuições do intervalo de tempo entre chegadas e do tempo de serviço, respectivamente, que incluem:

- M – Distribuição Exponencial Negativa para o intervalo de tempo entre chegadas sucessivas ou para o tempo de serviço;
- G – Distribuição não especificada (qualquer)
- D – Chegadas ou atendimentos determinísticos

Z - representa o número de servidores em paralelo

W - outras características do sistema, tais como comprimento da fila ilimitado ou população finita:

- em branco ou ∞ - modelo - base sem qualquer restrição adicional;
- K - comprimento da fila limitado, não podendo o número de elementos no sistema exceder K;
- N – população finita;

Modelo M/M/1 – Um único servidor

O que é?
M/M → Chegadas seguem distribuição de Poisson e tempos de serviço seguem uma distribuição **exponencial** (Markoviana).
1 → Existe **um único servidor** para atender a fila.

Exemplo 1 – Servidor Web simples
Um website pequeno recebe pedidos dos utilizadores. Cada pedido chega aleatoriamente (distribuição de Poisson). O servidor processa um pedido de cada vez. Se chegam pedidos enquanto o servidor está ocupado, eles **entram numa fila**.
Situação típica:
λ (taxa de chegada): 20 pedidos/minuto
μ (taxa de serviço): 30 pedidos/minuto

Modelo M/M/2 – Dois servidores paralelos

O que é?

M/M → Chegadas seguem distribuição de Poisson e tempos de serviço seguem uma distribuição **exponencial** (Markoviana).

2 → Existem **dois servidores paralelos** atendendo a mesma fila. Ambos têm a mesma taxa de serviço.

Exemplo 1 – Interface de Programação de Aplicações **com duas máquinas**

Uma IPA tem duas máquinas idênticas a processar pedidos. Se um pedido chega e uma máquina está livre, ele é atendido imediatamente. Se as duas estiverem ocupadas, o pedido fica na fila.

Situação típica:

λ = 40 pedidos/min

μ de cada servidor = 30 pedidos/min

Carga é distribuída entre as duas máquinas.

Exemplo: **Fila de atendimento num balcão de suporte técnico com capacidade limitada (k)**

Consideremos os Serviços de Informática da ESTG, que tratam de problemas nos computadores dos alunos/docentes /colaboradores/salas.

1. **M (Chegadas com distribuição de Poisson):** Os “clientes” aparecem aleatoriamente para pedir ajuda. A chegada não é previsível: às vezes chegam vários seguidos, outras vezes nenhum por longos períodos.

2. M (Tempos de serviço exponenciais): O técnico demora tempos variáveis para resolver o problema. Pode levar 2 minutos ou 20 minutos — depende da dificuldade.
3. 1 (Um único servidor): Há **um técnico** que atende um “cliente” de cada vez.
4. k (Capacidade máxima limitada) - O espaço é pequeno e só pode haver, por exemplo, **5 “clientes” no total**: 1 a ser atendido e 4 à espera; Se chegar o 6.º “cliente”, ele **não entra na fila.**

Estamos em presença de um modelo M/M/1/K=5

Exemplo: **Fila de atendimento num balcão de suporte técnico com população finita (N)**

Consideremos os Serviços de Informática da ESTG, que tratam de problemas nos computadores dos alunos/docentes /colaboradores/salas.

1. **M (Chegadas com distribuição de Poisson):** Os "clientes" aparecem aleatoriamente para pedir ajuda. A chegada não é previsível: às vezes chegam vários seguidos, outras vezes nenhum por longos períodos.

2. M (Tempos de serviço exponenciais): O técnico demora tempos variáveis para resolver o problema. Pode levar 2 minutos ou 20 minutos — depende da dificuldade.
3. 1 (Um único servidor): Há **um técnico** que atende um "cliente" de cada vez.
4. N (População finita) - A população de potenciais clientes é **limitada,** por exemplo: A escola tem **apenas 200 alunos**. Só esses 200 podem gerar pedidos.

Estamos em presença de um modelo M/M/1/N=200

Ver Formulário